2標本問題におけるブートストラップt検定

AUTHOR(S):

Wang, Jin Fang; Taguri, Masaaki

数理解析研究所講究録
916巻 1995年 1-11

2標本問題におけるブートストラップ t 検定

統数研　汪 金芳 (Jin Fang Wang)*
千葉大　田栗 正章 (Masaaki Taguri)†

概要

Possibilities of testing two means through nonparametric bootstrap approaches are discussed. The naive bootstrap by resampling the observed empirical distributions is useless in estimating null distributions. In simple random sampling, limited numerical investigations suggest resamples should be drawn from the original samples mixed, with or without proper transformations. In stratified sampling, the effects of location transformation on both size and power are also investigated through Monte Carlo simulations. Application is made to the historical two-sample problem of Darwin on crossed- and self-fertilized plant data.

Key Words: Alternative hypothesis; Location-aligned bootstrap; Mixed bootstrap; Monte Carlo simulation; Stratification.

1 Introduction

The main task of testing statistical hypotheses is to find null distributions of test statistics. For testing $H_0 : \mu_1 = \mu_2$, the equality of two means, based upon, for instance, the statistic

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2/m + S_Y^2/n}}, \tag{1}$$

requires finding $H_0(t)$, the distribution of T, when $\mu_1 = \mu_2 = \mu_0$ is assumed to be true. In (1), $\overline{X}, \overline{Y}, S_X^2$ and S_Y^2 are sample means and variances of the $i.i.d.$ random variables $X_1, \cdots, X_m$ from distribution $F(\mu_1)$ and $Y_1, \cdots, Y_m$ from distribution $G(\mu_2)$, respectively.

If F and G both are normal with common variances, then the null distribution of T is t_{m+n-2}, t-distribution with degrees of freedom $m+n-2$. The trouble is that if

*統計数理研究所 〒106 東京都港区南麻布 4-6-7　e-mail: wang@ism.ac.jp
†千葉大学理学部 〒263 千葉市稲毛区弥生町 1-33　e-mail: taguri@math.s.chiba-u.ac.jp

T is to be used and observations $x_1, \cdots, x_m$ and $y_1, \cdots, y_n$ display nonnormality how we should approximate the null distribution of T. The problem does not have an exact solution even in the normal case with heterogenous variances, which is refered to as the Behrens-Fisher problem in the literature. Bootstarp is a natural candidate in this kind of situations.

The naive bootstrap suggests estimating $H_0(t)$ by $\widehat{H}_0(t)$, the distribution of

$$T^* = \frac{\overline{X}^* - \overline{Y}^*}{\sqrt{S_X^{*2}/m + S_Y^{*2}/n}}, \tag{2}$$

where $\overline{X}^*, \overline{Y}^*, S_X^{*2}$ and S_Y^{*2} are sample means and variances of the empirical distributions $F_m(x)$ and $G_n(y)$, based on the observations $x_1, \cdots, x_m$ and $y_1, \cdots, y_n$, respectively. This algorithm failes catastrophically even in the simplest normal cases, see next section.

Section 2 starts the investigations from this simple normal cases: F and G are normal with possibly different variances. We propose the mixed bootstrap tests that resample the pooled original data with and without transformations. Section 3 applies the mixed bootstrap tests to historical two-sample problem of Darwin, yielding conclusions similar to those of classic analyses that the crossed plan may be superior to the self-fertilized. Section 4 considers more complicated situations when F and G are normal mixtures and samples are drawn from each subpopulation. Only location-aligned bootstrap is investigated in this case.

2 Mixed bootstrap tests

The failure of the naive bootstrap(Section 1) lies in the obvious fact that the bootstrap estimate $\widehat{H}_0(t)$ does not reflect the mechanism that $H_0(t)$ is produced under the constraint $\mu_1 = \mu_2$. One way of achieving this is to redefine the $\overline{X}^*, \overline{Y}^*, S_X^{*2}$ and S_Y^{*2} in (2) to be the sample means and variances of respective empirical distributions $\widehat{F}_m(x)$ and $\widehat{G}_n(y)$, putting mass $1/m$ and $1/n$ on $X_1^*, \cdots, X_m^*$ and $Y_1^*, \cdots, Y_n^*$, which are randomly drawn with replacement from

$$\{z_1, \cdots, z_{m+n}\} = \{x_1, \cdots, x_m;\ y_1, \cdots, y_n\}. \tag{3}$$

Let X and Y come from the null $F = N(\mu_0, \sigma^2)$, $G = N(\mu_0, \sigma^2)$. Let $m = n = 5$. Then $H_0(t) = t_8$ The first row of Table 1 shows the relative errors of the mixed bootstrap in approximating the lower and upper quantiles of t_8. As a comparision, the fourth row corresponding to the naive bootstrap test is also displyed.

The idea of mixing is not entirely new. Boos, Janssen and Veraverbeke(1989) discusses the pooled bootstrap tests for testing homogeneity of scales, which essentially uses the idea of mixing. An alternative way of forcing the two empiricals to have

the same mean is by location transformation. Efron and Tibshirani(1993, pp.224) suggests that the bootstrap samples be drawn from $\{x'_1, \cdots, x'_m\}$ and $\{y'_1, \cdots, y'_n\}$, where the x's and y's are location adjusted. They are defined as $x'_i = x_i - \overline{x} + \overline{z}$ and $y'_i = y_i - \overline{y} + \overline{z}$, where $\overline{x}$ and $\overline{y}$ are the respective sample means and $\overline{z}$ is the pooled mean. The fivth row of Thable 1 shows that this does not work well enough. The location-scale transformation, by redefining the x's and y's as $x'_i = (x_i - \overline{x})/S_x$ and $y'_i = (y_i - \overline{y})/S_x$, where S_x and S_y are the sample standard errors, improves the location transformation, but only mildly, see the last row of Table 1. The second and the third row of Table 1 compare the mixed bootstrap tests when location or location-scale transformation is applied before mixing, with simple mixed bootstrap test(first row). The results are essentially the same, with moderate improvements by including transformations.

Table 1 Errors in approximating the tails of the null distribtion by six bootstrap tests. The null distribtion of T under nomality with homogeneous scales, is $t_{5+5-2} = t_8$. The bootstrap tests use data from a "local alternative", $N(1, 1), N(0, 0.9^2)$.

	1%	2%	3%	4%	5%	95%	96%	97%	98%	99%
methods										
mixing	*.09*	*.04*	*.03*	*.03*	*.03*	*.01*	*.01*	*.02*	*.03*	*.05*
l-mixing	*.09*	*.02*	*.03*	*.03*	*.03*	*.02*	*.02*	*.01*	*.01*	*.03*
ls-mixing	*.10*	*.04*	*.03*	*.03*	*.03*	*.00*	*.00*	*.00*	*.01*	*.01*
naive	*.71*	*.87*	*.97*	*1.04*	*1.12*	*1.66*	*1.66*	*1.64*	*1.61*	*1.61*
location	*.31*	*.17*	*.12*	*.10*	*.10*	*.14*	*.17*	*.19*	*.23*	*.28*
location-scale	*.25*	*.15*	*.10*	*.09*	*.08*	*.03*	*.05*	*.07*	*.10*	*.15*

Notes: (1)The figures are relative errors defined by $|(w - \hat{w})/w|$, w and $\hat{w}$ stands for the true value and approximate value respectively; (2)mixing, l-mixing and ls-mixing stand for the mixed bootstrap, mixed bootstrap after location and location-scale transformation, respectively; naive, location and location-scale stand for the nonmixed bootstrap, nonmixed bootstrap with location and location-scale transformation, respectively; (3)The bootstrap values are averages during 100 repeated sampling, with each bootstrapped 200 times.

Similar features are observed from Table 2, where the null distribution of T based on $F = N(\mu_0, \sigma_1^2)$ and $F = N(\mu_0, \sigma_2^2)$, $\mu_0 = 1$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$ are approximated by 5,000 Monte Carlo trials($m = n = 5$).

Table 2 Errors in approximating the tails of the null distribtion by six bootstrap tests. The null distribtion of T is defined by assuming $X \sim N(1,1)$ and $Y \sim N(1,2^2)$, sample sizes $m = n = 5$. The bootstrap tests use data from a "local alternative", $N(1,\ 1), N(2,\ 2^2)$.

	1%	2%	3%	4%	5%	95%	96%	97%	98%	99%
methods										
mixing	*.08*	*.08*	*.05*	*.02*	*.03*	*.06*	*.06*	*.07*	*.06*	*.06*
l-mixing	*.09*	*.08*	*.04*	*.03*	*.02*	*.06*	*.06*	*.06*	*.06*	*.08*
ls-mixing	*.03*	*.07*	*.04*	*.01*	*.02*	*.05*	*.04*	*.04*	*.03*	*.02*
naive	*1.64*	*1.38*	*1.38*	*1.40*	*1.40*	*.79*	*.74*	*.67*	*.62*	*.46*
location	*.34*	*.19*	*.18*	*.18*	*.17*	*.23*	*.25*	*.28*	*.32*	*.48*
location-scale	*.11*	*.03*	*.04*	*.05*	*.03*	*.02*	*.01*	*.03*	*.07*	*.13*

Notes: (1)The figures are relative errors defined by $|(w - \hat{w})/w|$, w *and* $\hat{w}$ *stands for the true value and approximate value respectively; (2)mixing, l-mixing and ls-mixing stand for the mixed bootstrap, mixed bootstrap after location and location-scale transformation, respectively; naive, location and location-scale stand for the nonmixed bootstrap, nonmixed bootstrap with location and location-scale transformation, respectively; (3)The bootstrap values are averages during 100 repeated sampling, with each bootstrapped 200 times; (4)The null distribution is approximated by 5,000 Monte Carlo trials.*

3 Bootstrap Tests of Darwin's Zea Data

Table 3 shows data obtained by Darwin(1876), who investigated whether there exists superiority of the crossed plants over the self-fertilized. The data shown here concerns only zea, one out of the seven plants experimented by Darwin. The problem is to test the null hypothese $H_0 : \mu_X = \mu_Y$ against the alternative $H_1 : \mu_X > \mu_Y$, where μ_X and μ_Y represent the mean height of the crossed and the self-fertilized zea, respectively.

While the nonmixed bootstrap with location transformation applied to the zea data gives one-sided achieved significance level(a.s.l.) 0.043, the simple mixed bootstrap has a.s.l. 0.012, providing much stronger evidence against the null hypothesis that there is no difference between the crossed and self-fertilized zea. In mixing is done after location transformation, the a.s.l. decreases to 0.006, indicating even more discrepancy between the two kinds of zea. Location-scale transformation does not have much effect in the mixing case(with a.s.l. 0.011), but reduces the a.s.l. to 0.015 in the nonmixing case. All these a.s.l.'s are obtained using the same 2,000

bootstrap samples in the mixed and nonmixed cases, respectively. For comparision, the two-sided a.s.l.'s of some conventional nonparametric tests, namely the median, Wilcoxson and permutation test are 0.001, 0.003 and 0.024, respectively, see Takeuchi and Ohasi(1981) for details.

Table 3 Darwin's observations on the grouth rates of the crossed and self-fertilized zea. Numbers are expressed in eighths of an inch.

crossed	*188*	*96*	*168*	*176*	*153*	*172*	*177*	*163*
(X)	*146*	*173*	*186*	*168*	*177*	*184*	*96*	
self-fertilized	*139*	*163*	*160*	*160*	*147*	*149*	*149*	*122*
(Y)	*132*	*144*	*130*	*144*	*102*	*124*	*144*	

NOTES: The data can be found in Fisher(1960, pp.30), which are divided into bloks of sizes (3,3,5,4), corresponding to each pot. For example, (188, 96, 168) and (139, 163, 160) are from pot 1, and (168, 177, 184, 96) and (144, 102, 124, 144) are from pot 4, etc.

4 Stratified Sampling

In this section, we treat general stratified problems, assuming $F(x) = \sum_{l=1}^{L} w_l F_l(x)$ and $G(y) = \sum_{h=1}^{H} p_h G_h(y)$. Here each $F_l(x)\,(l = 1, \cdots, L)$ and $G_h(y)\,(h = 1, \cdots, H)$ represent the l- and h-th stratum distribution functions and $w_l\,(l = 1, \cdots, L)$ and $p_h\,(h = 1, \cdots, H)$ are the corresponding stratum weights, subject to $\sum_{l=1}^{L} w_l = \sum_{h=1}^{H} p_h = 1$. We only consider the location-aligned bootstrap test.

4.1 The model

Suppose data $\{x_{l1}, \cdots, x_{lm_l}\}\,(l = 1, \cdots, L)$ and $\{y_{h1}, \cdots, y_{hn_h}\}\,(h = 1, \cdots, H)$, are observed from each stratum F_l and G_h, respectively. Let $\sum m_l = m$, $\sum n_h = n$. The sample means $\overline{x}_l = \sum x_{li}/m_l$ and $\overline{y}_h = \sum y_{hi}/n_h$ are unbiased estimates for each stratum mean μ_{lX} and μ_{hY}, which, combined together, form unbiased estimates $\overline{x}^s = \sum w_l \overline{x}_{lX}$ and $\overline{y}^s = \sum p_h \overline{y}_{hY}$ for the total mean μ_X and μ_Y respectively. Hereafter we only consider proportional allocation, i.e. $m_l = w_l m$ and $n_h = p_h n$.

Let $\widehat{\sigma}^2_{lX}$ be the usual unbiased version of sample variances of each stratum variance σ^2_{lX}. Wakimoto(1971) proved that

$$ {}_{st}\widehat{\sigma}^2_X = \sum_{l=1}^{L} w_l \widehat{\sigma}^2_{lX} + \sum_{l=1}^{L} w_l (\overline{x}_l - \overline{x}^s)^2 - \sum_{l=1}^{L} w_l (1 - w_l) \widehat{\sigma}^2_{lX}/m_l $$

is unbiased for the total variance σ^2_X. The unbiased estimator ${}_{st}\widehat{\sigma}^2_Y$ for σ^2_Y is similarly defined.

The stratified version of the usual t-statistic becomes

$$T_{st} = \frac{\overline{X}^s - \overline{Y}^s}{\sqrt{{}_{st}\widehat{\sigma}_X^2/m + {}_{st}\widehat{\sigma}_Y^2/n}}, \tag{4}$$

based on which we are to test $H_0 : \mu_X = \mu_Y$ against the alternative $H_1 : \mu_X > \mu_Y$.

To perform a test based on T_{st}, is to find, or to make a good approximation of $Q(F_N, G_N) = Prob(T_{st} \leq t)$, the null distribution function. To emphasize, we have deliberately changed our notation using F_N and G_N in place of F and G to represent the two distributions under null hypothesis, i.e. $\mu_X = \mu_Y$.

Let $\widehat{F}_l$ be the empirical distribution function of the l-th stratum of F putting mass $1/m_l$ on each atom $x_{li}\,(i = 1, \cdots, m_l)$, and $\widehat{G}_h$ similarly defined. Define $\widehat{F} = \sum_{l=1}^{L} w_l \widehat{F}_l$ and $\widehat{G} = \sum_{h=1}^{H} p_h \widehat{G}_h$. The naive bootstrap draws *i.i.d.* stratified samples with replacement from $\widehat{F}$ and $\widehat{G}$, exactly in the same way as the original stratified samples are drawn from F and G, which is as useless as in the simple random case.

4.2 Location-aligned bootstrap test

The location-aligned bootstrap test constitutes the following steps.

(1) Let $\overline{z} = (m\overline{x}^s + n\overline{y}^s)/(m+n)$. Define the pseudo-observations for $l = 1, \cdots, L$ and $h = 1, \cdots, H$ by

$$\begin{aligned} x_{li}^{+} &= x_{li} - \overline{x}_l + \overline{z}/Lw_l \quad (i = 1, \cdots, m_l), \\ y_{hi}^{+} &= y_{hi} - \overline{y}_h + \overline{z}/Hp_h \quad (i = 1, \cdots, n_h). \end{aligned}$$

(2) Define pseudo-empirical distribution functions

$$\begin{aligned} \widehat{F}_N &= \sum_{l=1}^{L} w_l \widehat{F}_{lN}, \\ \widehat{G}_N &= \sum_{h=1}^{H} p_h \widehat{G}_{hN}, \end{aligned}$$

where $\widehat{F}_{lN}$ is the empirical distribution function putting mass $1/m_l$ on each atom $x_{li}^{+}\,(i = 1, \cdots, m_l)$, and $\widehat{G}_{hN}$ is similar.

(3) Define the bootstrap estimate of $Q(F_N, G_N)$ by $Q(\widehat{F}_N, \widehat{G}_N)$, which is further approximated by Monte Carlo means

$$\frac{1}{B}\sharp\{T_{st}^{b*} \leq t\},$$

where T_{st}^{b*} is the version of T_{st}, based on the b-th stratified samples from $\widehat{F}_N$ and $\widehat{G}_N$, $\sharp$ stands for the number of the event within $\{\ \}$ being true and B is the number of the whole procedure replicated. Several remarks are pertinent.

Remark 1 To reflect the null hypothesis, namely two distributions sharing the same mean, one has no apparent reason for adjusting the scales. However, if the

null hypothesis *does* pose some restictions on the second order moments, as in the case of approximating the t-distribibition discussed in Section 2, proper adjustments upto that order may to be prefered. Empirical variances may be adjusted in the stratified case as following

$$x^o_{li} = [(x_{li} - \overline{x}_l)/s_{lx}](1/L\sqrt{w_l}) + 1/w_l L\sqrt{m} \quad (i = 1, \cdots, m_l),$$
$$y^o_{hi} = [(y_{hi} - \overline{y}_h)/s_{hy}](1/H\sqrt{p_h}) + 1/p_h H\sqrt{n} \quad (i = 1, \cdots, n_h).$$

This transformation is exact when $(m+1)/(n+1) = H/L$, which is satisfied, for example, when $m = n$ and $L = H$. We will not consider location-scale transformation in our Monte Carlo studies.

Remark 2 Confidence intervals based on asymptotically pivotal quantities tend to be long-shaped. A 95% bootstrap-t confidence interval for the difference between the crossed and self-fertilized zea in Example 1 is (1.8, 32.3), compared with the so-called nonparametric ABC interval(Efron and Tibshirani 1993), (5.5, 29.4). Welch's solution gives (3.1, 44.5), Fisher's fiducial interval is (2.7, 39.1), compared with (13, 39) which is based on Wilcoxson test(Takeuchi and Ohasi 1981, pp.51–89). One consequence of this is that "t-type" tests may tend to have lower power.

5 Monte Carlo Studies

Assume that the data do come from distributions satifying the null hypothesis and t_0 is the observed value of the stratified t-statistic. The achieved significance level, or p-value, $Prob(T_{st} > t_0)$ under null hypothesis depends solely on t_0, which has the uniform distribution on $(0, 1)$ if t_0 is randomly observed from the null hypothesis. We shall evaluate our bootstrap tests by checking the size, power and testing the uniformity of the p-values.

Now the hypothesis $H_0 : \mu_X = \mu_Y$ is to be tested against the alternative $H_1 : \mu_X > \mu_Y$, based on the location-aligned bootstrap test described in the previous section. For simplicity, we assume F and G in the following simulations to be mixtures of two normal populations, and $w_l = p_l (l = 1, 2)$. With no loss of generality, we fix $\mu_Y = 1$ and varify the following quantities: coefficient of variations, $C_X = \sigma_X/\mu_X$, $C_Y = \sigma_Y/\mu_Y$; ratio of (sub-)variances, $S_X = \sigma^2_{2X}/\sigma^2_{1X}$, $S_Y = \sigma^2_{2Y}/\sigma^2_{1Y}$; and the effect of stratifications, $\rho^2_X = \sum^2_{l=1}(\mu_{1X} - \mu_{2X})^2/\sigma^2_X$, $\rho^2_Y = \sum^2_{l=1}(\mu_{1Y} - \mu_{2Y})^2/\sigma^2_Y$. To fully specify F and G we need one more condition, which is designed so that the test is supposed to have approximate power (0.2, 0.4, 0.6, 0.8). This constraint is derived by approximating the bootstrap distribution of T^*_{st} by the limit $N(0, \sigma^2_A)$ of T_{st} under H_0 and $N(\delta, \sigma^2_A)$ under H_1, where $\sigma^2_A = [(1 - \rho^2_X)\sigma^2_X/m + (1 - \rho^2_Y)\sigma^2_Y/n]/(\sigma^2_X/m + \sigma^2_Y/n)$, and $\delta = (\mu_X - \mu_Y)/\sqrt{\sigma^2_X/m + \sigma^2_Y/n}$.

Table 4 shows relative good behaviour of the locatio-aligned bootstrap test, when $m = n = 10$, and the effects of stratification($\rho^2_X = \rho^2_Y = 0.3$) is relatively small.

The lower half of this table displays the results of the same bootstrap test when the stratified samples are treated as simple random samples, losses of power in the later case are observed.

Table 4 10%-level location-aligned bootstrap tests applied to normal mixtures. The sample sizes are $m = n = 10$, effects of stratification, $\rho_X^2 = \rho_Y^2 = 0.3$. The bootstrap tests approximately achieve the norminal level(0.1), and have reasonable power. Lower half of the table corresponds to the stratified samples misused as simple random samples.

weight(w_1)	*case*	*null*(p)	alt_1	alt_2	alt_3	alt_4
0.3	*I*	*.086(13.1)*	*.164*	*.384*	*.524*	*.758*
	II	*.118(13.5)*	*.150*	*.354*	*.512*	*.718*
0.5	*I*	*.080(6.4)*	*.146*	*.312*	*.486*	*.722*
	II	*.094(4.8)*	*.162*	*.278*	*.460*	*.682*
0.7	*I*	*.090(9.1)*	*.138*	*.368*	*.540*	*.768*
	II	*.086(12.0)*	*.186*	*.322*	*.520*	*.714*
0.3	*I*	*.070(13.0)*	*.134*	*.322*	*.458*	*.710*
	II	*.078(8.7)*	*.096*	*.276*	*.464*	*.630*
0.5	*I*	*.072(14.7)*	*.130*	*.290*	*.496*	*.712*
	II	*.054(17.8)*	*.132*	*.240*	*.414*	*.612*
0.7	*I*	*.060(22.4)*	*.104*	*.312*	*.494*	*.736*
	II	*.044(24.8)*	*.096*	*.202*	*.370*	*.598*

NOTES: (1) Case I and II correspond to the parameter layout $S_X = S_Y = 0.5$, $C_X = C_Y = 0.3$ *and* $S_X = S_Y = 1$, $C_X = 0.3$, $C_Y = 0.8$, *respectively; (2) values in* () *are p-values of* χ_9^2 *to test the uniformity of the a.s.l. in approximating the null distributions,* (90%, 95%)*-percentiles of* χ_9^2 *being (14.7, 16.9); (3) The simulations are based on 500 repeated sampling, each bootstrapped 500 times.*

Location-aligned bootstrap tests are however quite sensitive to the effects of stratification (ρ_X^2, ρ_Y^2), and to the balance of samples, as can be seen from Table 5. To improve, ideas like mixing may be incorporated, we leave the experiments as our future task.

Table 5 10%-level location-aligned bootstrap tests applied to two normal mixtures. Lower half of the table corresponds to the stratified samples misused as simple random samples.

$\rho_X^2 = \rho_Y^2$	$weight(w_1)$	$null(p)$	alt_1	alt_2	alt_3	alt_4
0.3	*0.3*	*.026(34.0)*	*.068*	*.168*	*.278*	*.504*
	0.5	*.026(32.4)*	*.064*	*.138*	*.288*	*.494*
	0.7	*.028(32.8)*	*.060*	*.204*	*.312*	*.538*
0.8	*0.3*	*.000(55.6)*	*.000*	*.002*	*.026*	*.042*
	0.5	*.000(55.6)*	*.000*	*.000*	*.006*	*.030*
	0.7	*.000(55.6)*	*.000*	*.004*	*.012*	*.038*
0.3	*0.3*	*.036(26.3)*	*.084*	*.234*	*.370*	*.570*
	0.5	*.042(20.8)*	*.082*	*.174*	*.332*	*.554*
	0.7	*.022(36.0)*	*.078*	*.202*	*.332*	*.570*
0.8	*0.3*	*.000(55.6)*	*.000*	*.002*	*.028*	*.068*
	0.5	*.000(55.6)*	*.000*	*.002*	*.010*	*.048*
	0.7	*.000(55.6)*	*.000*	*.002*	*.008*	*.030*

NOTES: (1) Sample sizes $m = 20$, $n = 10$, *other parameters:* $S_X^2 = S_Y^2 = 0.5$, $C_X = C_Y = 0.3$; *(2) values in () are p-values of* χ_9^2 *to test the uniformity of the a.s.l. in approximating the null distributions, (99%, 99.5%)-percentiles of* χ_9^2 *are (21.7, 23.6); (3) The simulations are based on 500 repeated sampling, each bootstrapped 500 times.*

5.1 Darwin's example revisited

Darwin planted his plants in different pots. He was carful to make the conditions in each pot as near as possible. But we still hope the information on pot can be utilized in the inference. We put data in Pot 1 and 2 in stratum 1 and the rest as stratum 2, since the mixed pots having near means. The results are summerized in Table 6, which are quite consistant with traditional tests. Stratification does seem to provide more information.

Table 6. Bootstrap tests of the difference on growth rates between the crossed- and self-fertilized zea. The figures are two-sided achieved significance levels, obtained from 200 bootstrap samples. Other classical parametric and nonparametric tests are also shown(Takeuchi and Ohasi, 1981). Among these tests, the median test has least a.s.l., the corresponding 95% confidence interval is (19, 45), which is "significantly" short. See Remark 2 od Section 4.2.

Method	**Type of Transformation**	**asl**
	no transformation	*.505*
stratified sampling	*location*	*.000(.005)*
	location-scale	*.015(.005)*
simple	*no transformation*	*.545*
randon	*location*	*.025(.010)*
sampling	*location-scale*	*.005(.005)*
$N(\mu_x, \sigma^2)$, $N(\mu_y, \sigma^2)$		*.02*
one-sample t(d.f.=n-1)		*.05*
Wilcoxson		*.003*
permutation		*.024*
median		*.0014*

NOTES: (1)Strata are formed by mixing Pot 1 and 2, and Pot 2 and 3;(2) $N(\mu_x, \sigma^2)$, $N(\mu_y, \sigma^2)$ *stands for the method based the normal assumptions with homogeneous variances; one-sample t(d.f.=n-1) for Fisher's one-sample t-test by properly pairing the data(see Fisher, 1960); Wilcoxson for Wilcoxson test; permutation for permutation test; and median for median test; (3)Figures in brackets are obtained from mixing the transformed data in simple random sampling, but only mixing the transformed data within each population in stratified situations.*

6 Discussions

Bootstrap tests are statistical procedures for seeking information about models in one class(the null class), conditional on information about a different class(the alternative class). In a strict sense, we never have direct information of the first class, but always observe instead information of the latter class. "Validity" of the transformations of the *mal*information into the information we want, obviously depends upon the structural relationship between the two classes.

参考文献

[1] Boos, D., Janssen, P. and Veraverbeke, N.(1989). Resampling from centered data in the two-sample problerm, *Journal of Statistical Planning and Inference,* 21, 327-345.

[2] Darwin, C.(1876), *The effects of cross- and self-fertilisation in the vegetable kingdom,* London: John Murray.

[3] Efron, B. and Tibshirani, R.(1993),*An Introduction to the Bootstrap*: Chapman and Hall.

[4] Fisher, R.A.(1960), *The Design of Experiments*(7th ed.), Edinburgh: Oliver and Boyd.

[5] Takeuchi, K. and Ohasi, Y.(1981), *Toketekisuisoku–2 hyohonmondai*(in Japanese), Tokyo: Nihonhyoronsya.

[6] Wakimoto, K.(1971), Stratified random sampling (I), Estimation of the population variance. *Ann. Inst. Statist. Math.*, 23, 233-252.